Estimating infectious disease incidence: validity of capture-recapture analysis and truncate models for incomplete countdata.

van Hest, R.; Grant, A.; Smit, H.F.E.; Story, A.; Richardus, J.H.

published in
Epidemiology and Infection
2007

DOI (link to publisher)
10.1017/S0950268807008254

document version
Publisher's PDF, also known as Version of record

Epidemiol. Infect. (2008), **136**, 14–22. © 2007 Cambridge University Press
doi:10.1017/S0950268807008254 Printed in the United Kingdom

Estimating infectious diseases incidence: validity of capture–recapture analysis and truncated models for incomplete count data

N. A. H. VAN HEST[1,2]*, A. D. GRANT[3], F. SMIT[4,5], A. STORY[6] AND J. H. RICHARDUS[1,2]

[1] *Division of Infectious Disease Control, Municipal Public Health Service Rotterdam Area, Rotterdam, The Netherlands*
[2] *Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam, The Netherlands*
[3] *Statistics, Modelling and Bioinformatics Department, Centre for Infections, Health Protection Agency, London, UK*
[4] *Trimbos Institute, Netherlands Institute of Mental Health and Addiction, Utrecht, The Netherlands*
[5] *Department of Clinical Psychology, Vrije Universiteit, Amsterdam, The Netherlands*
[6] *Respiratory Diseases Department, Centre for Infections, Health Protection Agency, London, UK*

(*Accepted 11 February 2007; first published online 12 March 2007*)

SUMMARY

Capture–recapture analysis has been used to evaluate infectious disease surveillance. Violation of the underlying assumptions can jeopardize the validity of the capture–recapture estimates and a tool is needed for cross-validation. We re-examined 19 datasets of log-linear model capture–recapture studies on infectious disease incidence using three truncated models for incomplete count data as alternative population estimators. The truncated models yield comparable estimates to independent log-linear capture–recapture models and to parsimonious log-linear models when the number of patients is limited, or the ratio between patients registered once and twice is between 0·5 and 1·5. Compared to saturated log-linear models the truncated models produce considerably lower and often more plausible estimates. We conclude that for estimating infectious disease incidence independent and parsimonious three-source log-linear capture–recapture models are preferable but truncated models can be used as a heuristic tool to identify possible failure in log-linear models, especially when saturated log-linear models are selected.

INTRODUCTION

Surveillance of infectious diseases is an essential part of public health. Mandatory notification is one of the mechanisms to carry out such surveillance but under-notification has been widely reported. For meaningful interpretation of the number of patients with infectious diseases the completeness of notification should be estimated. This can be done through a statistical technique called capture–recapture analysis. Based on certain assumptions, capture–recapture methods use information on the overlap of linked disease registers to estimate the number of patients unknown to all registers and thus the estimated total number of patients [1]. Completeness of notification can then be assessed relative to the estimated total number of patients. In biomedical sciences capture–recapture analysis is frequently used for estimating the number

* Author for correspondence: N. A. H. van Hest, M.D., M.Sc., Tuberculosis Control Physician/Epidemiologist, Division of Infectious Disease Control, Municipal Public Health Service Rotterdam Area, PO Box 70032, 3000 LP Rotterdam, The Netherlands.
(Email: vanhestr@ggd.rotterdam.nl)

of accidents and injuries [2] and patients with mostly chronic diseases such as congenital deformities [3], insulin-dependent diabetes mellitus [4], cancer [5], neurological conditions [6] or rheumatological diseases [7]. Less frequently it has been used for evaluating infectious disease surveillance, especially when record-linkage is based on more than two registers.

The validity of capture–recapture estimates depends on possible violations of the underlying assumptions: cases can be uniquely identified (i.e. registers have a perfect positive predictive value), perfect record-linkage (i.e. no misclassification of records), a closed population (i.e. no immigration or emigration in the time period studied) and a homogeneous population [i.e. no subgroups with markedly different (re)capture probabilities]. In two-source capture–recapture methods one must also assume independence between registers [i.e. the probability of being observed in one register is not affected by being (or not being) observed in the other registers]. In the three-source capture–recapture approach pairwise dependencies, i.e. dependencies between two registers, can be identified and accounted for in a log-linear model [1, 8–11]. The three-way (highest-order) interaction, however, i.e. dependency between all three registers, cannot be incorporated in the model and its absence must be assumed.

In epidemiological studies violation to some degree of most of the underlying capture–recapture assumptions is unavoidable. This and other limitations of capture–recapture analysis are described elsewhere in more detail [10, 12–19]. Infectious diseases carry an elevated risk that some capture–recapture analysis assumptions are violated. Especially absence of dependence between the available registers, including three-way interaction, and heterogeneity among the patients cannot be excluded and should be expected. Consequently, the validity of two-source and three-source capture–recapture studies requires critical scrutiny.

Sometimes, it becomes evident that a capture–recapture model breaks down and produces erratic results. While performing three-source log-linear model capture–recapture studies on the completeness of notification of tuberculosis in The Netherlands [20] and England we were confronted with unexpected and unrealistic estimates of tuberculosis incidence, despite using well-described procedures for finding the best log-linear model [21]. In this context, solely relying on three-source capture–recapture analysis without any cross-validation seems to be inappropriate. We suggest that three-source capture–recapture analyses should be complemented by alternative methods to arrive at, and cross-validate, estimates of population size. Alternative models related to capture–recapture analysis have been described and offer the opportunity to cross-validate outcomes. The aim of this study is to re-examine the data of published and current three-source log-linear model capture–recapture studies on infectious disease incidence with various truncated models for incomplete count data and describe the apparent agreement or discrepancy of the estimates.

METHODS

Data sources

Data sources used were 19 datasets in 16 published or current three-source log-linear model capture–recapture studies on infectious disease incidence known to us.

Truncated population estimators

The data sources were re-examined with three alternative population estimators: a truncated binomial model, a truncated Poisson mixture model (Zelterman) [22] and a truncated Poisson model (Chao) [23, 24]. Out of the many possible methods we have chosen this combination of truncated models because they have been described as an alternative to capture–recapture methods [10, 25], can be used on the same data that is needed for the three-source log-linear model and are easy to apply [26, 27].

In epidemiology, truncated estimators are usually applied to frequency counts of observations of individuals in a single data source [28]. They aim to estimate the number of unobserved persons (falling in the zero-frequency class) based upon information on the number of times a person has been observed. Technically, one assumes a specific truncated distribution of the observed data, e.g. Poisson or binomial, and then extrapolates from the observed series to the unobserved number of people never seen [10]. Observed frequency distributions may not be strictly Poisson and to relax this assumption Zelterman based his model on a Poisson mixture distribution, allegedly allowing greater flexibility and applicability on real-life data [28]. Conceptual aspects of the Zelterman and Chao models have been discussed in some detail elsewhere [27, 29–31]. The simple truncated estimators do not need statistical packages. In the social

sciences truncated models have been employed to estimate the size of hidden populations such as criminals [26, 32], illegal residents [33] and illicit drug users and homeless persons [27–29, 34]. To our knowledge, truncated estimators have not been used before to estimate the number of infectious disease patients.

As with capture–recapture analysis, the validity of the estimates of truncated models depends on the possible violation of the underlying assumptions. These assumptions are similar to the capture–recapture assumptions described earlier but in addition equiprobability (i.e. equal ascertainment probabilities of all registers) should be assumed when using multiple sources [10]. Some truncated models are arguably more robust to population heterogeneity because they are partly based upon the lower frequency classes, and the people seen rarely are assumed to have a greater resemblance with the people never seen. This relative insensitivity to violation of the homogeneity assumption of some truncated estimators is supported mathematically and through simulation studies but these can occasionally underestimate the true population size in the presence of heterogeneity [22, 29].

Frequency counts

It is possible to extract frequency counts for the truncated models from multiple-source capture–recapture data, allowing us to use the reported data from the log-linear studies for the truncated models. The ratio between the number of patients registered once (f_1) and registered twice (f_2) plays an important role in the truncated models. When '1' represents being known to a register and '0' represents being unknown to a register, and three linked registers are used, frequency count f_1 is the sum of the cells 100, 010 and 001 in the $2 \times 2 \times 2$ contingency table and frequency count f_2 corresponds to the sum of the cells 110, 101 and 011. Similarly, patients observed in all three registers, f_3, are denoted as 111. For all 19 datasets the number of patients in these seven cells are shown later. We use the f_1/f_2 ratio to examine a possible relationship between this ratio and the performance of truncated models *vis-à-vis* the log-linear models.

RESULTS

Table 1 (available in the online version of the paper) shows the various three-source log-linear model capture–recapture studies of infectious disease incidence and completeness of notification with the number of patients observed and their frequency counts, the objective of the study, the data sources used and the selected log-linear model. The studies involved eight infectious diseases and were performed at the local, regional or national level. One study collected data over a 4-months period, the other studies over 1- to 5-year periods. The observed number of patients varied from 33 to 28 678. Notification, laboratory and hospital registers were the most conventional data sources used. The distribution of the patients over three linked registers in the various three-source capture–recapture studies of infectious diseases is shown in Table 2 (available online).

The log-linear and truncated model estimates with their respective confidence and prediction intervals are shown in Table 3 (available online), as well as the f_1/f_2 ratio among the observed patients and the coefficient of variation of the data source probabilities (see Discussion). The capture–recapture studies varied in estimated number of patients from 46 to 42 969. A second truncated Poisson estimator, Chao's bias-corrected homogeneity model,

$$\text{est}(N) = \text{obs}(N) + [(f_1^{\,2} - f_1)/(2(f_2 - 1))],$$

gave similar estimates as Chao's heterogeneity model [35]. A second truncated binomial estimator,

$$\text{est}(N) = \text{obs}(N)/[1 - (1/(1 + f_2/f_1))^3],$$

gave similar estimates as the truncated binomial model used (data not shown).

f_1/f_2 ratio

On the basis of the f_1/f_2 ratio the studies can be divided in four categories:

(*a*) $f_1/f_2 < 0{\cdot}5$ (dataset 7). In this study all estimates were similar but the number of observed patients was small.

(*b*) $0{\cdot}5 < f_1/f_2 < 1{\cdot}5$ (datasets 1, 2, 6, 8–10, 13a, 13b, 14, 16a, 16b). In these studies the truncated binomial model and Zelterman's model gave similar results as the independent (without interactions) or parsimonious log-linear model while Chao's model estimates were slightly higher. When a saturated log-linear model (with all two-way interactions) was selected the truncated estimates were considerably lower than the log-linear model estimates.

(*c*) $1{\cdot}5 < f_1/f_2 < 3{\cdot}5$ (datasets 5, 11, 12, 15). In the first study the results of all truncated models were similar to the parsimonious log-linear model estimate but the number of observed patients was small. In the second study the estimates of Zelterman's and Chao's truncated models were lower but within the 95% confidence interval (CI) of the parsimonious log-linear model estimate while the truncated binomial model estimate was considerably lower. In the third study all truncated model estimates were considerably lower than the saturated log-linear model estimate, the truncated binomial estimate again being lowest. In the fourth study all truncated model estimates were lower than the saturated log-linear model estimate but fell within the broad 95% CI, the truncated binomial model estimates again lowest.

(*d*) $f_1/f_2 > 3{\cdot}5$ (datasets 3a, 3b, 4). In all studies the truncated model estimates were considerably higher than the parsimonious log-linear model estimates, especially the Zelterman and Chao models.

Selected log-linear model

On the basis of the selected log-linear model the studies can be divided in three categories:

(*a*) Independent log-linear model (datasets 1, 2). In these studies the truncated models produce similar estimates as the log-linear model.

(*b*) Parsimonious log-linear model (datasets 3–11). In the 11 studies with a parsimonious log-linear model selected three observations can be made:

- In the three studies with $f_1 \gg f_2$ (datasets 3, 4) the truncated binomial model estimates a higher number of patients than the log-linear model while the truncated Poisson and Poisson mixture models estimate a considerably higher number of patients.
- In the three studies (datasets 5–7) with a small number of observed patients the estimates of the log-linear model and truncated Poisson, Poisson mixture and binomial models are comparable.
- In the studies with the f_1/f_2 ratio between 0·5 and 1·5 (datasets 8–10, 13b) the truncated model estimates are similar to the log-linear model but the Chao models can be relatively higher and in one study the truncated Poisson mixture estimate was relatively low.

(*c*) Saturated log-linear model (datasets 12–16, apart from 13b). In all but one of the studies with a saturated model selected (datasets 12, 13a, 14, 16) the truncated models gave considerable lower and mutually comparable estimates.

DISCUSSION

Main findings

In three-source log-linear model capture–recapture studies of infectious disease incidence with an independent log-linear model selected, truncated models yield comparable estimates. The truncated models also give similar results when parsimonious log-linear models are selected and the number of patients is limited or the f_1/f_2 ratio is between 0·5 and 1·5. When $f_1 \gg f_2$ truncated models give considerable higher estimates than parsimonious log-linear models. Compared to saturated log-linear models the truncated models produce considerably lower and often more plausible estimates.

Capture–recapture analysis and chronic diseases

For human diseases capture–recapture analysis has predominantly been applied to estimate the prevalence, incidence or completeness of registers of specific groups of diseases, often diseases with a chronic character as mentioned earlier. Apparently the characteristics of most of these diseases, their patients and their registers best fulfil criteria for feasibility of capture–recapture studies as well as validity of the underlying assumptions. Perhaps with the exemption of some neurological and rheumatological conditions the case-definition is probably unambiguous and uniform over the various registers. Arguably, for these categories of diseases sufficient registers are available and possible relationships between these registers, e.g. clinical registers, laboratory registers, health insurance registers or patient support and advocacy group registers, be they positive or negative, could be avoided by source selection or source merging or accounted for in a log-linear model, thus limiting violation of the independent registers assumption. The permanent character of most of these conditions can reduce violation of the closed population assumption.

Capture–recapture analysis and infectious diseases

For infectious diseases the number of available registers for record-linkage, usually notification-, laboratory- or hospital-based registers, is often limited and (strong) positive interaction between these registers should be expected as a result of the characteristics of infectious disease diagnosis and treatment, and public health regulations. Infectious disease control and surveillance is often organized around close collaboration between clinicians, microbiologists and public health professionals, such as infectious disease and tuberculosis physicians and nurses. Only two of the 19 datasets studied selected the independent log-linear model and 11 datasets selected parsimonious log-linear models incorporating one or two pairwise dependencies. However, six datasets selected the saturated log-linear model, i.e. including all two-way interactions and assuming absence of the three-way interaction [16, 36]. Our studies of tuberculosis incidence in England and, before correction for suggested imperfect record-linkage and remaining false-positive hospital cases, in The Netherlands both selected a saturated model, resulting in unexpectedly and unrealistically high estimates of the number of tuberculosis patients. The two previous three-source log-linear model capture–recapture studies of tuberculosis incidence resulted in a parsimonious model and both produced plausible estimates within the range of prior expectations [37, 38]. According to Hook & Regal, if the saturated model is selected by any criterion the investigator should be particularly cautious about using the associated outcome [10]. At the time of our studies on tuberculosis incidence all but one of the published three-source log-linear capture–recapture studies of infectious incidence used independent or parsimonious log-linear models (studies 1–11). The one published study selecting a saturated log-linear model (study 12) gave a much higher estimate ($n = 1314$) of the number of hepatitis A patients in an outbreak in Taiwan than later established by serology results ($n = 545$) [19]. Recently a three-source log-linear model capture–recapture study of meningococcal disease incidence also selected a saturated log-linear model and resulted in relatively high estimates (study 16) [39]. Perhaps confidence in the validity of capture–recapture results may reflect publication bias in favour of apparently successful capture–recapture studies [40]. The unexpectedly high estimates of the saturated log-linear model capture–recapture studies do not result from violation of the 'absent three-way interaction' assumption. In the case of infectious disease registers, existing three-way interaction is almost certainly positive, causing a capture–recapture estimate biased downwards [39]. The reason for the high estimates must, therefore, be violation of (a combination of) the other underlying assumptions. After correction for possible false-positive records and possible imperfect record-linkage the capture–recapture studies on tuberculosis and meningococcal disease in The Netherlands (studies 13 and 16) produced much lower and lower estimates, respectively. Compared to an initial saturated log-linear model, a covariate log-linear capture–recapture model, reducing violation of the homogeneity assumption, also resulted in a much lower estimate of 886 (95% CI 827–1022) Legionnaires' disease patients in The Netherlands (study 15).

Truncated estimators and infectious diseases

Infectious disease studies where an independent log-linear model was selected produce estimates very similar with the truncated models, which can be partly explained by the independent register assumption underlying the truncated models when applied to three registers. That truncated estimators perform well when data are sparse is demonstrated in studies 5, 6 and 7 as the estimates of the log-linear and the various truncated models are similar. The truncated models also give similar results as the log-linear models when $0{\cdot}5 < f_1/f_2 < 1{\cdot}5$ but give considerably higher estimates when $f_1 \gg f_2$. In the case of saturated log-linear models (studies 12–16), with unexpectedly high estimates of infectious disease incidence, the lower truncated model estimates are more plausible but are they also preferable? We have two arguments to support the view they might be:

(1) In study 12 the saturated log-linear model estimated 1314 patients with hepatitis A infection in an outbreak in Taiwan while the truncated models estimate between 500 and 600 patients. The National Quarantine Service of Taiwan, on the basis of serology tests, later concluded that the true number of infected persons was about 545, making this one of the few capture–recapture datasets where later a true number of patients was established [19].
(2) A saturated log-linear model in dataset 13a gave an implausible estimate of 2053 (95% CI 1871–2443) tuberculosis patients in The Netherlands in

1998, while truncated models estimated between 1600 and 1675 patients. The implausible estimate caused the investigators to have a critical look at the data again and make further corrections for probable imperfect record-linkage and possible remaining false-positive records in the hospital register. The parsimonious log-linear model of dataset 13b fitted the adjusted data well and gave an estimate of 1547 (95% CI 1513–1600) tuberculosis patients and corresponding truncated model estimates. The initial truncated model estimates came relatively close to the final log-linear model estimate.

The equiprobability and number of data source assumptions

The truncated binomial model assumes that all sources have the same probability of capturing a case. In addition the truncated Poisson model assumes an infinite number of sources, although in our data the number of sources was limited to three. On this argument the truncated binomial model for three data sources is a more realistic alternative estimator. However, any departure from equiprobability results in an estimation error, which analytically is overestimation (see Appendix). Realistic estimates of this error can be obtained from the data. In Table 3 the last column shows the coefficients of variation, a measure of variability in the coverages of the three data sources for each study. This is calculated as the standard deviation divided by the mean from the three quantities N_1 (number of cases known on source 1), N_2 (number of cases known on source 2) and N_3 (number of cases known on source 3). We demonstrate the possible effect of violation of the equiprobability assumption by studies 4 and 11. For study 4, which has a high coefficient of variation (0·86), if the sources were truly independent, the number of unobserved cases would be 702, calculated by fitting the log-linear model with main effects only. Our truncated binomial estimator gives 1325 cases, nearly twice as large. For study 11, with a low coefficient of variation (0·06), independence implies that there are 155 unobserved cases, while the truncated binomial estimate is 212, an overestimation by about 30%. Studies 3 and 4 indicate that the high f_1/f_2 ratios result from violation of the equiprobability assumption, producing overestimates by the truncated models.

Two-source validation

Any three-source study can be used to test two-source estimation by treating one source as though it were a complete list of cases and extract a complete 2×2 table. We demonstrate this for two studies, numbers 4 and 11, which we chose above for their coefficients of variation and took register 3 as the complete set. Validation was by comparing the Petersen estimator ($N_{10}\, N_{01}/N_{11}$) [1] and the truncated binomial estimator, which for two lists is $f_1^2/(4f_2)$, on the 2×2 table with the known 'unlisted' number. For study 4 there were 451 'unlisted' cases, i.e. on neither of registers 1 and 2. The Petersen estimator is 37 and the truncated binomial estimator 42. The two estimators are similar because registers 1 and 2 have approximately equal coverage but both are far short of the true figure (Zelterman and Chao models estimates are 79 and 84, respectively). For study 11 there were 161 'unlisted' cases and the two estimators were 57 and 64. Again the estimators agree but are short of the true figure. Now the Zelterman and Chao model estimates are 107 and 130, respectively, and perform slightly better. However, we had some hesitation in extracting 2×2 tables from three-source capture-recapture data, more specifically from capture–recapture studies on infectious disease incidence. As explained earlier, (positive) interdependencies between the three conventional registers used for such studies should be expected. Extracting 2×2 tables ignores possible conditional dependence confounding the results thus obtained. The log-linear model in study 4 included one interaction term for pairwise dependencies and the log-linear model in study 11 included two such interaction terms, which may explain the underestimation in the Petersen and truncated estimators. We therefore also validated the two studies with independent log-linear models (studies 1 and 2). We took register 2 as the complete set for study 1 and register 3 as the complete set for study 2. For study 1 there were 73 'unlisted' cases. The Petersen estimator, 43, is a little low, but the truncated binomial estimator, at 201, is too high (Zelterman and Chao model estimates are 397 and 401, respectively). The discrepant (over)estimate by the truncated models can be explained by the different coverages of registers 1 and 3, i.e. violation of the equiprobability assumption. In study 2 the coefficient of variation was low and the coverage of registers 1 and 2 similar. For study 2 there were 22 'unlisted' cases. The Petersen estimator and the truncated binomial estimator are

both 25 and similar to the known 'unlisted' number, explained by almost absent violation of both the independent sources and equiprobability assumptions. The Zelterman and Chao model estimates are 43 and 51, respectively and the discrepancy with the truncated binomial model estimate can be explained by violation of the 'infinite number of sources' assumption.

Alternative models

As an alternative to log-linear capture–recapture models a structural source model has been proposed [36]. Whereas log-linear models only partly identify and incorporate dependencies between registers, the structural source model models potential interdependencies of the registers and heterogeneity of the population, partly based on prior knowledge, and estimates the probabilities of conditions that produce these interactions between the registers. However, the published data of the capture–recapture studies were insufficient to re-examine these studies with a structural source model.

CONCLUSION

We have indicated conditions where estimates of infectious disease incidence from log-linear models are similar or dissimilar to alternative truncated models for incomplete count data. Our results suggest that for estimating infectious disease incidence and completeness of notification independent and parsimonious three-source log-linear capture–recapture models are preferable. When saturated models are selected as best-fitting model and the estimates are unexpectedly high and seem implausible, first, the data should be re-examined with truncated models as a heuristic tool, in the absence of a gold standard, to identify possible failure in the saturated log-linear model when the truncated models produce a lower estimated number of infectious disease patients. Second, in case of such discrepancy between the log-linear and the truncated model estimates, the data should be re-examined for possible violation of the underlying capture–recapture assumptions, such as imperfect record-linkage, false-positive records or heterogeneity, corrected and the capture–recapture analysis repeated on the corrected data. When after repeated capture–recapture analysis the discrepancy between the log-linear and the truncated model estimates remains or no violation of the underlying assumptions can be identified, the investigator should be cautious about using the associated outcome [10]. Using truncated model estimates as an early alert could prevent flawed capture–recapture estimates finding their way into the scientific literature. The role of the f_1/f_2 ratio in the agreement or disagreement between three-source log-linear capture–recapture and truncated model estimates for the number of infectious disease patients, especially when a parsimonious log-linear model is selected, should be the subject of further mathematical or statistical studies.

APPENDIX

Equations for the truncated population estimators

Truncated binomial model:

$$\text{est}(N) = \text{obs}(N) + (f_1)^2/3f_2.$$

Truncated Poisson mixture model:

$$\text{est}(N) = \text{obs}(N)/[1 - \exp(-2f_2/f_1)].$$

Truncated Poisson heterogeneity model:

$$\text{est}(N) = \text{obs}(N) + (f_1)^2/2f_2.$$

Equiprobability

If the truncated binomial model is true, i.e. if the sources are independent and equiprobable with probability of capturing any case $= p$, our estimator $(f_1)^2/(3f_2)$ is correct in the sense that the expected number of unlisted cases is given by

$$\mathbf{E}f_0 = Nq^3 = \frac{(\mathbf{E}f_1)^2}{3\mathbf{E}f_2}. \quad (1)$$

If we introduce a small departure from equiprobability so that the list probabilities are $(p-h, p, p+h)$ instead of (p, p, p), the estimation error can be defined as

$$g(h,p) = \frac{(\mathbf{E}f_1)^2}{3\mathbf{E}f_2} - \mathbf{E}f_0. \quad (2)$$

Differentiating with respect to h, we find that

$$g(0,p) = \frac{\partial g}{\partial h}(0,p) = 0; \quad \frac{\partial^2 g}{\partial h^2}(0,p) = \frac{2N(1-p)}{3p^2}, \quad (3)$$

so that we overestimate, at least for small h. The same happens if we consider an asymmetrical departure

$(p-h, p, p)$. In that case,

$$g(0,p)=\frac{\partial g}{\partial h}(0,p)=0; \quad \frac{\partial^2 g}{\partial h^2}(0,p)=\frac{2N(1-p)}{9p^2}, \qquad (4)$$

and there is again an overestimate.

DECLARATION OF INTEREST

None.

NOTE

Supplementary information accompanies this paper on the Journal's website (http://journals.cambridge.org).

REFERENCES

1. **International Working Group for Disease Monitoring and Forecasting.** Capture-recapture and multiple-record estimation I: history and theoretical development. *American Journal of Epidemiology* 1995; **142**: 1047–1058.
2. **LaPorte RE, *et al.*** Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. *American Journal of Epidemiology* 1995; 15; **142**: 1069–1077.
3. **Orton H, Richard R, Miller L.** Using active medical record review and capture-recapture methods to investigate the prevalence of Down Syndrome among live-born infants in Colorado. *Teratology* 2001; **64**: S14–19.
4. **EURODIAB ACE Study Group.** Variation and trends in incidence of childhood diabetes in Europe. *Lancet* 2000; **355**: 873–876.
5. **McClish D, Penberthy L.** Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. *Medical Care* 2004; **42**: 1111–1116.
6. **Tilling K, Sterne JA, Wolfe CD.** Estimation of the incidence of stroke using a capture-recapture model including covariates. *International Journal of Epidemiology* 2001; **30**: 1351–1359.
7. **Mahr A, *et al.*** Prevalences of polyarteritis nodosa, microscopic polyangiitis, Wegener's granulomatosis, and Churg-Strauss syndrome in a French urban multiethnic population in 2000: a capture-recapture estimate. *Arthritis & Rheumatism* 2004; **51**: 92–99.
8. **Fienberg SE.** The multiple-recapture census for closed populations and the 2k incomplete contingency table. *Biometrika* 1972; **59**: 591–603.
9. **Bishop YM, Fienberg SE, Holland PW.** *Discrete Multivariate Analysis*. Cambridge: MIT Press, 1975.
10. **Hook EB, Regal RR.** Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**: 243–263.
11. **International Working Group for Disease Monitoring and Forecasting.** Capture-recapture and multiple-record estimation II: Applications in human diseases. *American Journal of Epidemiology* 1995; **142**: 1059–1068.
12. **Desenclos JC, Hubert B.** Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* 1994; **23**: 1322–1323.
13. **Brenner H.** Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; **6**: 42–48.
14. **Cormack RM.** Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology* 1999; **52**: 909–914.
15. **Papoz L, Balkau B, Lellouch J.** Case counting in epidemiology: limitations of methods based on multiple data sources. *International Journal of Epidemiology* 1999; **25**: 474–478.
16. **Hook EB, Regal RR.** Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**: 771–779.
17. **Jarvis SN, *et al.*** Children are not goldfish-mark-recapture techniques and their application to injury data. *Injury Prevention* 2000; **6**: 46–50.
18. **Tilling K.** Capture-recapture methods-useful or misleading? *International Journal of Epidemiology* 2001; **30**: 12–14.
19. **Chao A, *et al.*** The applications of capture-recapture models to epidemiological data. *Statistics in Medicine* 2001; **20**: 3123–3157.
20. **Van Hest NAH, *et al.*** Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiology and Infection*. Published online: 7 December 2006. doi: 10.1017/S0950268806007540.
21. **Van Hest NAH, Smit F, Verhave JP.** Improving malaria notification in the Netherlands: results from a capture-recapture study. *Epidemiology and Infection* 2002; **129**: 371–377.
22. **Zelterman D.** Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference* 1988; **18**: 225–237.
23. **Chao A.** Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; **43**: 783–791.
24. **Chao A.** Estimating animal abundance with capture frequency data. *Journal of Wildlife Management* 1988; **52**: 295–300.
25. **Hook EB, Regal RR.** Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *American Journal of Epidemiology* 1982; **116**: 168–176.
26. **Van der Heijden PG, Cruyff MJ, Van Houwelingen H.** Estimating the size of a criminal population from police registrations using the truncated Poisson regression model. *Statistica Neerlandica* 2003; **57**: 289–304.

27. **Smit F, Toet J, Van der Heijden PG.** Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data. In: Hay G, McKeganey N, Birks E, eds. *Final report EMCDDA project Methodological Pilot Study of Local Level Prevalence Estimates*. Lisbon: EMCDDA, 1997, pp. 47–66.
28. **Bohning D, *et al.*** Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology* 2004; **19**: 1075–1083.
29. **Wilson RM, Collins MF.** Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; **79**: 543–553.
30. **Smit F, Reinking D, Reijerse M.** Estimating the number of people eligible for health service use. *Evaluation and Program Planning* 2002; **25**: 101–105.
31. **Hay G, Smit F.** Estimating the number of hard drug users from needle-exchange data. *Addiction Research and Theory* 2003; **11**: 235–243.
32. **Rossmo DK, Routledge R.** Estimating the size of criminal populations. *Journal of Quantitative Criminology* 1990; **6**: 293–314.
33. **Bustami R, *et al.*** Point and interval estimation of the population size using the truncated Poisson regression model. In: Klein B, Korsholm L, eds. *New Trends in Statistical Modelling. Proceedings of the 16th International Workshop on Statistical Modelling*. Odense: University of Southern Denmark, 2001, pp. 87–94.
34. **Hser YI.** Population estimation of illicit drug users in Los Angeles County. *Journal of Drug Issues* 1993; **23**: 323–334.
35. **Chao A.** Estimating population size for sparse data in capture-recapture experiments. *Biometrics* 1989; **45**: 427–438.
36. **Regal RR, Hook EB.** Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Statistics in Medicine* 1998; **17**: 69–74.
37. **Tocque K, *et al.*** Capture recapture as a method of determining the completeness of tuberculosis notifications. *Communicable Diseases and Public Health* 2001; **4**: 141–143.
38. **Baussano I, *et al.*** Undetected burden of tuberculosis in a low-prevalence area. *International Journal of Tuberculosis and Lung Disease* 2006; **10**: 415–421.
39. **De Greeff SC, *et al.*** Underreporting of meningococcal disease incidence in the Netherlands: results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *European Journal of Epidemiology* 2006; **21**: 315–21.
40. **Hay G.** The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; **46**: 515–520.